Report

Name: Shashank Vinod Chardeve & Chinmay H R

Title: Fraudulent Claim Detection

Introduction:

The insurance business faces a serious problem with insurance fraud, which raises expenses for both honest policyholders and insurers and results in large financial losses. By examining a large dataset with several parameters pertaining to policies, clients, and occurrences, this research seeks to create an effective and precise machine learning-based method for identifying fraudulent insurance claims. In order to help insurance companies decrease fraudulent payouts and increase operational efficiency, the project aims to detect suspicious claims with high sensitivity and precision by utilizing models like logistic regression and random forest in conjunction with meticulous data preparation and feature engineering.

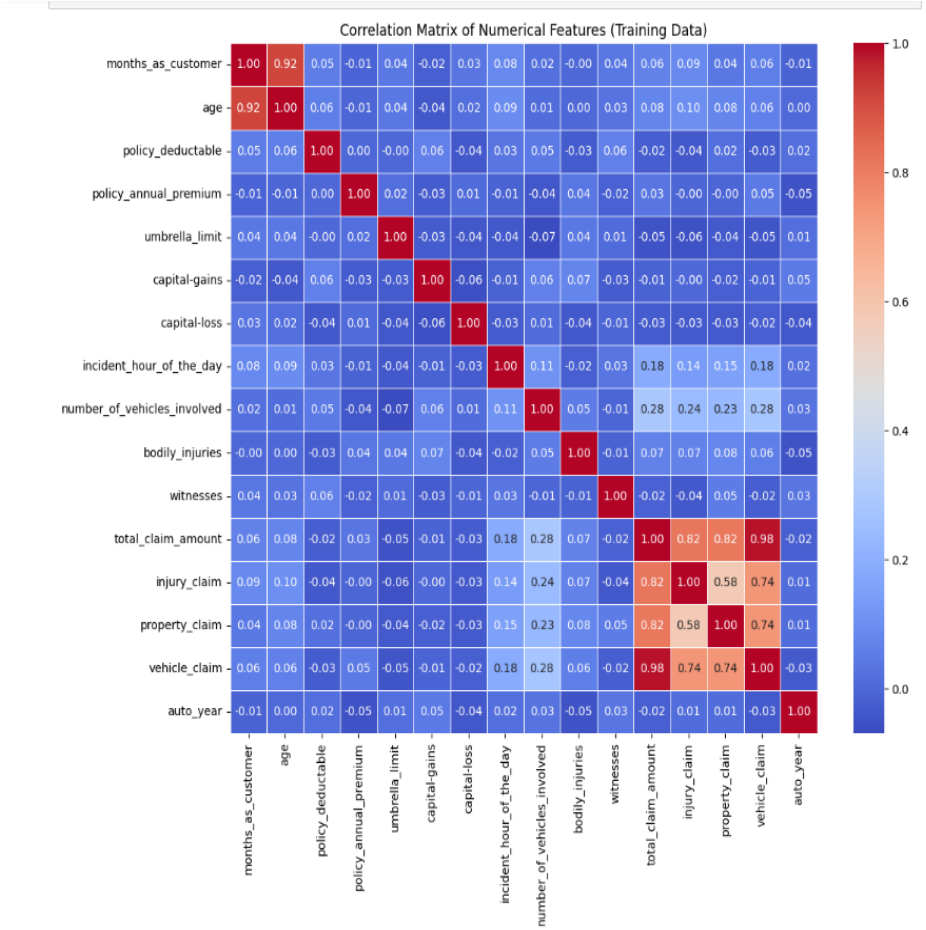
2) Dataset Overview:

The 1,000 insurance claim records in the dataset utilized for this research have 40 different features covering policy details, client demographics, incident details, and claim information. Policy numbers, dates, incident kinds, vehicle and driver characteristics, and monetary claim amounts are among the categories and numerical variables it contains. The dataset is appropriate for supervised machine learning problems centered on fraud detection since it is labeled with a target variable that indicates whether each claim is authentic or fraudulent. This extensive dataset allows for in-depth research and model training to successfully distinguish between authentic and fraudulent claims.

3) Exploratory Data Analysis:

- The goal of the training data's Exploratory Data Analysis (EDA) was to comprehend the distributions of both numerical and categorical features, their relationships, and how they relate to the target variable, according to fraud. The distributional characteristics of important variables, including months as a customer, age, policy deductible, annual premium, incident hour, number of vehicles involved, bodily injuries, witnesses, claim amounts, and vehicle year, were discovered by univariate analysis of numerical features. The prevalence of high-value claims that could affect model sensitivity was shown by histograms with kernel density estimates, which showed that the majority of numerical variables showed skewness, with some distributions (such as total claim amount and injury claim) being right-skewed. This foundational knowledge was essential for handling outliers or additional modifications.

- Relationships between variables were found through correlation analysis using a heatmap of numerical features. This revealed some moderate correlations, particularly between claim-related features such as injury claim and total claim amount, as well as minor dependencies between policy deductible and annual premium. These observations are useful for identifying possible multicollinearity and guiding the process of feature selection or dimension reduction.



- A class balance analysis revealed an unbalanced target variable with fewer false claims than true ones, highlighting the necessity of using methods like resampling to correct for this imbalance while training the model.

- For categorical variables, bivariate analysis evaluated the probability of fraud in various groups. Features such as insurance state, policy coverage limitations, insured education level, occupation, hobbies, incident type, collision type, incident severity, authorities contacted, and vehicle make/model showed notable differences in fraud rates. For instance, there was a larger chance of fraud in occupations like executive-managerial and lower fraud rates in vocations like handler-cleaner. The chance of fraud was significantly higher in multi-vehicle collisions than in incident types such as vehicle theft and parked cars. These trends aid in distinguishing between categorical variables that have a significant impact on fraud risk and those that have no effect.

4) Feature Engineering:

- This insurance fraud detection project's feature engineering methodology included a number of essential phases to address data issues and increase model accuracy. First, the training data was subjected to a RandomOverSampler technique in order to address the class imbalance where false claims were underrepresented. By randomly replicating minority class samples, this technique balanced the dataset, improving the model's capacity to detect fraud and reducing bias towards the majority class.

- The difference in days between the event date and the policy bind date was then introduced as a new feature ("policy_incident_duration") in order to capture significant temporal information. Finding timing patterns that could indicate fraud is the goal of this feature. In order to simplify the dataset and eliminate superfluous information, redundant columns—such as the original date fields from which this new feature was derived—were eliminated.

- To further increase model generalization and decrease noise, low-frequency categories were grouped under a common label "Other" to refine category columns with numerous unique or sparse values. In addition to managing uncertain data by substituting "?" with "Unknown" in collision type, property damage, and police report availability, this covered features such insured occupation, hobbies, car make, and model.

- All categorical features were then transformed into numerical format appropriate for machine learning methods by creating dummy variables. Data consistency was maintained by taking care to guarantee that, following dummy encoding, the columns in the training and validation datasets were aligned. For both datasets, the target variable was also encoded in binary format.

- In order to ensure that all numerical inputs were standardized to a common scale, numerical characteristics were finally scaled using the StandardScaler. This stabilizes and enhances learning performance by preventing characteristics with wider ranges from unduly affecting the models.

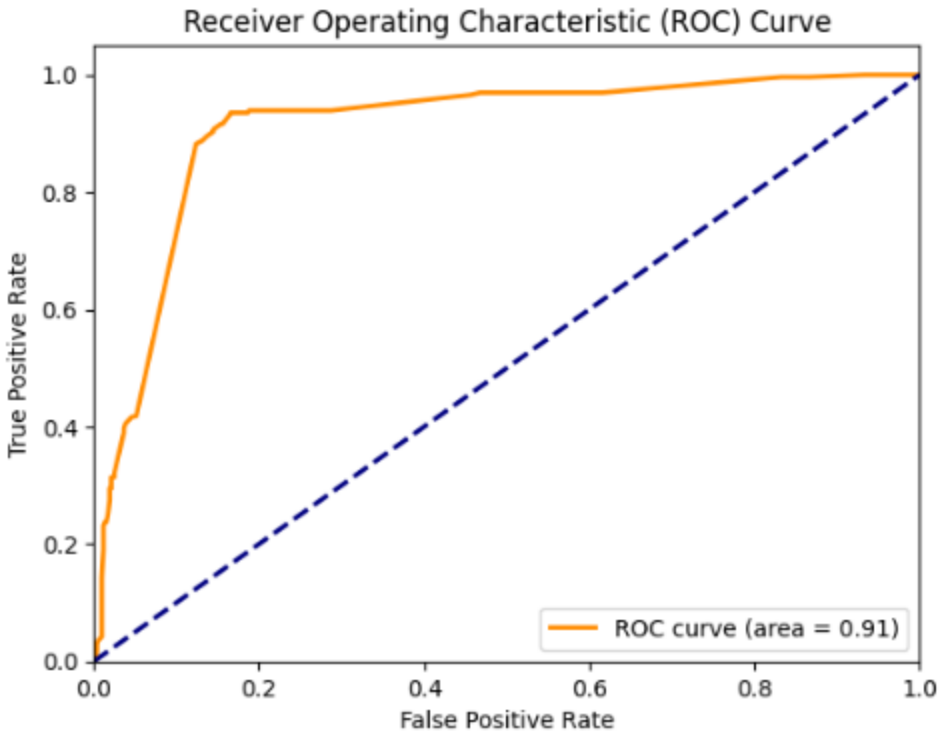
- All things considered, these feature engineering procedures methodically addressed class imbalance, improved information richness, decreased noise, and readied the dataset for the construction of reliable and efficient machine learning models.

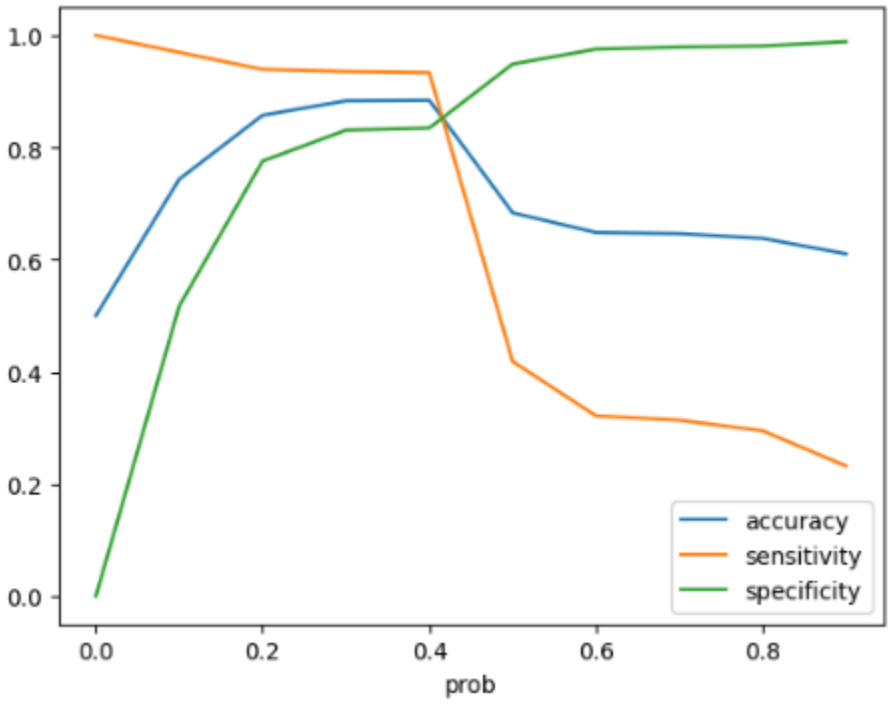
5) Model Selection:

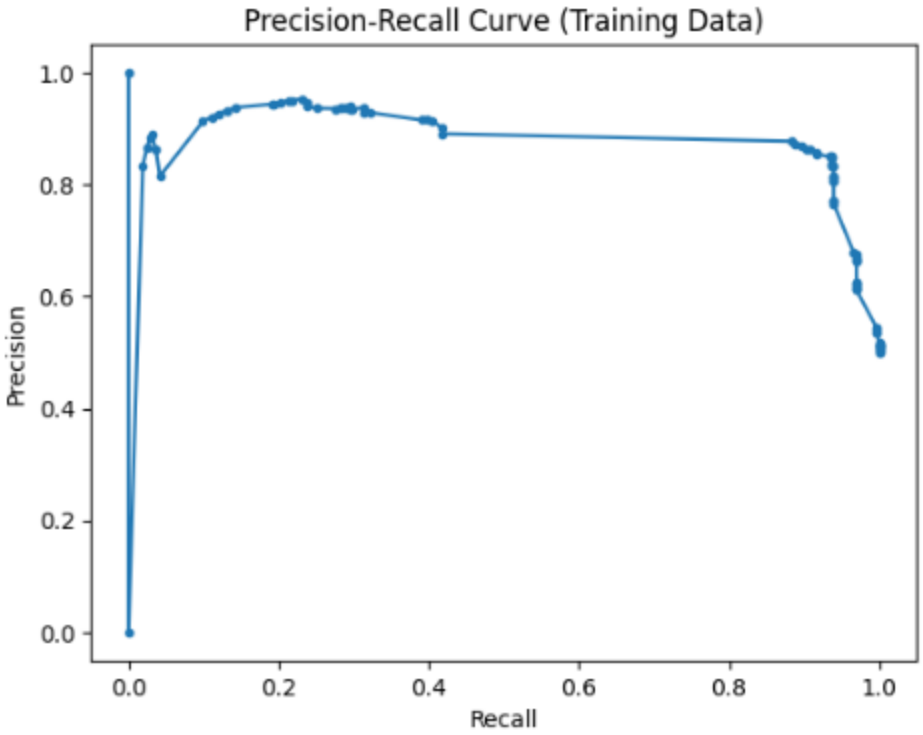
The model building phase of the insurance fraud detection project involved developing and optimizing two key machine learning models: Logistic Regression and Random Forest, each offering complementary strengths for fraud classification.

(A) Logistic Regression Model:

Model coefficients and p-values demonstrated the significance of variables such as "insured_hobbies_chess" and "incident_severity_Minor Damage," while Variance Inflation Factors (VIFs) confirmed the lack of troublesome multicollinearity. The initial model evaluation on training data revealed moderate accuracy (~68%), high precision (0.89) but relatively low sensitivity (0.42), indicating difficulty in capturing all fraudulent cases at the default classification cutoff of 0.5. A detailed cutoff optimization using ROC and precision-recall analysis determined that 0.2 was the ideal probability threshold, significantly enhancing sensitivity to 0.94 and obtaining an overall accuracy of roughly 86%, with balanced specificity (0.78) and F1-score (0.87). This careful adjustment improved the model's ability to detect fraud without producing an excessive number of false positives. For the Logistic Regression model, Recursive Feature Elimination with Cross-Validation (RFECV) was used to identify the most relevant subset of features, yielding 15 significant predictors such as specific insured hobbies, occupations, incident severity levels, and auto models. These features were utilized to build the logistic regression model with statsmodels, which allows for thorough statistical analysis.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
ROC curve (area = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensitivity
specificity
prob
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve (Training Data)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

(B) The Random Forest Model:

The Random Forest model was designed to supplement logistic regression with a powerful ensemble learning technique. The most influential predictors were identified by feature importance analysis, which included customer tenure, age, claim amounts, event timing, and chosen category variables. The baseline random forest model obtained 100% accuracy on the training data, demonstrating great learning capabilities. Cross-validation revealed substantial generalization, with mean accuracy above 92% and no significant overfitting. Further performance improvement was achieved through systematic hyperparameter tweaking using GridSearchCV, which adjusted parameters such as tree depth (max_depth=15), number of trees (n_estimators=300), and leaf/branch splitting criteria. The tweaked random forest model retained perfect training accuracy and good cross-validation scores, indicating an optimal bias-variance balance. During training, sensitivity, specificity, accuracy, recall, and F1-scores all exceeded 1.0, demonstrating the model's potential as a highly accurate fraud classifier.

(C) Hyperparameter Tuning:

In summary, logistic regression delivers interpretable feature-level insights as well as a configurable classification threshold for balancing fraud detection sensitivity and specificity. Random forest is a highly adaptable, high-performing black-box model that can capture complicated nonlinear interactions, particularly after hyperparameter adjustment. Together, these models offer complementary benefits, allowing for effective fraud mitigation techniques through a mix of explainability and predictive capability.

6) Prediction and Model Evaluation:

- During the prediction and model evaluation phase, fraud predictions were generated based on training data, and model performance was assessed using different metrics to verify reliability and efficacy.

- Initial predictions for the logistic regression model were made with a standard probability threshold of 0.5, resulting in a moderate accuracy of about 68%. The confusion matrix revealed a high true negative count but a relatively low true positive count, yielding a sensitivity of 0.42 and specificity of 0.95. While precision was good at 0.89, inadequate sensitivity indicated that many false claims were overlooked. To solve this, a detailed cutoff optimization approach was carried out, which involved evaluating model performance across a range of probability thresholds from 0.0 to 0.9. This study used ROC curves to investigate the trade-off between true positive rate and false positive rate, as well as precision-recall curves to balance fraud detection accuracy versus false alarms. The ideal cutoff was 0.2, which increased sensitivity to 0.94 while keeping a reasonable specificity of 0.78 and accuracy of 86%. This adjustment demonstrated how threshold tweaking can improve fraud detection by emphasizing recall without significantly reducing precision. The final F1-score of 0.87 demonstrated a great balance of precision and recall, confirming the model's classification capacity under the increased cutoff.

- The random forest model's predictions on the training data initially achieved 100% accuracy, as seen in a confusion matrix with no false positives or false negatives. Evaluation parameters such as sensitivity, specificity, accuracy, recall, and F1-score all reached the desired value of 1.0, indicating a perfect separation of fraudulent and valid claims in the training set. Five-fold cross-validation was used to ensure the model's generalizability and prevent overfitting, yielding consistently high validation accuracies of more than 92%. The random forest's robustness was increased further by hyperparameter tuning using grid search, which optimized values for tree depth, number of estimators, and minimum sample splits. The tweaked model maintained perfect accuracy on training data and strong validation performance, demonstrating its potential for use in fraud detection settings.

- In conclusion, both models demonstrated strong predictive capabilities, with logistic regression offering interpretable and tunable thresholding for fraud sensitivity, and random forest delivering high accuracy and robustness through ensemble learning and hyperparameter optimization. Together, these evaluations confirm the models' effectiveness in accurately detecting insurance fraud, providing multiple tools for enhancing fraudulent claim identification in practical settings.

7) Conclusion:

This project successfully demonstrated the application of machine learning techniques to detect fraudulent insurance claims, a significant problem causing financial losses in the insurance industry. By employing logistic regression with feature selection and threshold optimization, alongside a robust random forest model with hyperparameter tuning, the project achieved high predictive accuracy and balanced detection metrics. The models effectively addressed challenges of class imbalance and complex feature interactions through careful data preparation and feature engineering. Logistic regression provided interpretability with statistical insights, while random forest delivered superior accuracy and robustness. Together, these models form a complementary solution for automating fraud detection, reducing false positives, and improving operational efficiency. Future work could enhance adaptability by incorporating new data sources and advanced algorithms, ensuring continued protection against evolving fraud tactics.